

Linear Size Distance Preservers

Greg Bodwin*

Stanford University
Department of Computer Science

Abstract

The famous shortest path tree lemma states that, for any node s in a graph $G = (V, E)$, there is a subgraph on $O(n)$ edges that preserves all distances between node pairs in the set $\{s\} \times V$. A very basic question in distance sketching research, with applications to other problems in the field, is to categorize when *else* graphs admit sparse subgraphs that preserve distances between a set P of p node pairs, where P has some different structure than $\{s\} \times V$ or possibly no guaranteed structure at all. Trivial lower bounds of a path or a clique show that such a subgraph will need $\Omega(n + p)$ edges in the worst case. The question is then to determine when these trivial lower bounds are sharp; that is, when do graphs have *linear size distance preservers* on $O(n + p)$ edges?

In this paper, we make the first new progress on this fundamental question in over ten years. We show:

1. All G, P has a distance preserver on $O(n)$ edges whenever $p = O(n^{1/3})$, even if G is directed and/or weighted. These are the first nontrivial preservers of size $O(n)$ known for directed graphs.
2. All G, P has a distance preserver on $O(p)$ edges whenever $p = \Omega\left(\frac{n^2}{\text{RS}(n)}\right)$, and G is undirected and unweighted. Here, $\text{RS}(n)$ is the Ruzsa-Szemerédi function from combinatoric graph theory. These are the first nontrivial preservers of size $O(p)$ known in any setting.
3. To preserve distances within a subset of s nodes in a graph, $\omega(s^2)$ edges are sometimes needed when $s = o\left(\frac{n^{2/3}}{2^{\Theta(\sqrt{\log n \cdot \log \log n})}}\right)$ even if G is undirected and unweighted. For weighted graphs, the range of this lower bound improves to $s = o(n^{2/3})$. This result reflects a polynomial improvement over lower bounds given by Coppersmith and Elkin (SODA '05).

An interesting technical contribution in this paper is a new method for “lazily” breaking ties between equally short paths in a graph, which allows us to draw our new connections between distance sketching and the Ruzsa-Szemerédi problem.

*gbodwin@cs.stanford.edu

1 Introduction

We begin by recalling the famous Shortest Path Tree Lemma:

Lemma (Shortest Path Trees). *Given a node u in an n -node graph $G = (V, E)$, there is a subgraph $H \subseteq G$ on $O(n)$ edges such that the distances of G and H agree on all node pairs in $\{u\} \times V$.*

Needless to say, this lemma holds a central place in theoretical computer science; it is an indispensable structural result in graph theory, and the computation of shortest path trees is one of our most basic algorithmic problems (e.g. Dijkstra’s Algorithm). This has led researchers to consider the following natural generalization: how many edges are needed in H if we want the distances of G and H to agree on a set of node pairs P , which does not necessarily have the structure $P = \{u\} \times V$?

These subgraphs, which generalize shortest path trees, are called *pairwise distance preservers*:

Definition (Pairwise Distance Preservers – [19], following [14]). *Given a graph $G = (V, E)$ and a set of node pairs $P \subseteq V \times V$, a subgraph $H \subseteq G$ is a pairwise distance preserver of G, P if the distances of G and H agree on all node pairs in P .*

The goal of research on distance preservers is simply to understand what the worst-case density of H must be, given the number of nodes in G and pairs in P . A preliminary trivial lower bound is that $\Omega(n)$ edges are needed to preserve just a single pairwise distance when the underlying graph is a path, and that $\Omega(p)$ edges are needed to preserve any p pairwise distances when the underlying graph is a clique. Thus, the best upper bound we can hope for in this problem is $O(n + p)$ edges in a preserver. Note that this bound is upper bound is realized in the case of shortest path trees; that is, $O(n + p)$ edges always suffice when $P = \{s\} \times V$. A very fundamental question, which is an even more direct generalization of shortest path trees, is then to simply categorize when *else*, if ever, we are guaranteed the existence of “linear size distance preservers” on $O(n + p)$ edges. This question is the main focus of this paper.

1.1 History

Pairwise distance preservers were first studied by Coppersmith and Elkin [19], following a closely related notion of “ D -Preservers” by Bollobás, Coppersmith, and Elkin [14]. Perhaps surprisingly, Coppersmith and Elkin showed that nontrivial linear size distance preservers do indeed exist: in any undirected graph, one can preserve any $p = O(n^{1/2})$ pairwise distances using a subgraph on just $O(n)$ edges. This unexpected fact has found several interesting applications: for example, Elkin and Pettie [26] used it to build the first linear-size $\log n$ stretch path reporting distance oracle, Bodwin and Vassilevska W. [13] have used it to build the current most accurate additive spanner of linear size, Pettie [38] used them as an ingredient in state-of-the-art constructions for mixed spanners, and they were employed by Elkin, Filtser, and Neiman [24] to build terminal subgraph spanners.

Coppersmith and Elkin [19] also showed some lower bounds on the existence of linear size preservers. They showed that $\omega(n + p)$ edges are needed to preserve $\omega(n^{1/2}) = p = \frac{n^2}{2^{\Omega(\sqrt{\log n \cdot \log \log n})}}$ pairwise distances in undirected unweighted graphs; when the underlying graph can be weighted, the range of this lower bound improves to $\omega(n^{1/2}) = p = o(n^2)$ (i.e. no further linear size preservers are possible). These lower bounds were recently used in constructions by Abboud and Bodwin [1, 2] for spanner lower bounds; in particular, these distance preserver lower bounds were shown to imply the surprising fact that graphs cannot be compressed into $n^{4/3-\varepsilon}$ bits while preserving distance

		Upper Bound	Lower Bound
Unwtd.	Undir.	$O(np^{1/3} + n^{2/3}p^{2/3})$ and $O(n + n^{1/2}p)$ O(n) when $p = O(n^{1/2})$	$\Omega(n^{2d/(d^2+1)}p^{d(d-1)/(d^2+1)}), d \in \mathbb{N}$ $\omega(n + p)$ when $\omega(n^{1/2}) = p = o(n^{2-o(1)})$
	Dir.	$O(np^{1/2})$	
Wtd.	Undir.	$O(n + n^{1/2}p)$ and $O(np^{1/2})$ O(n) when $p = O(n^{1/2})$	$\Omega(n^{2/3}p^{2/3})$ $\omega(n + p)$ when $\omega(n^{1/2}) = p = o(n^2)$
	Dir.	$O(np^{1/2})$	

Figure 1: State of the art upper and lower bounds for pairwise distance preservers, for directed/undirected and weighted/unweighted graphs. The first upper bound for undirected unweighted graphs is due to [13], and the remaining bounds in this chart are all due to [19]. The hidden $n^{o(1)}$ factor in the unweighted lower bound has the form $2^{O(\sqrt{\log n \cdot \log \log n})}$.

information up to any constant additive error, thus resolving a long-standing open question about the quality of additive distance compression.

While applications of distance preserver bounds continue to mount (see also e.g. [14, 12, 5]), no further progress has been made on the existence/nonexistence of linear size distance preservers since the original paper by Coppersmith and Elkin in 2005. Several major gaps in our knowledge of distance preservers remain. For example, it is remarkable that **no** separations between the various graph settings are currently known: it is conceivable that the worst-case sparsity of distance preservers in directed weighted graphs is exactly the same as their worst-case sparsity in undirected unweighted graphs. The difference between the directed and undirected settings is particularly striking here. *No* interesting linear size distance preservers are known for directed graphs; the only known upper bound of any sort is that $O(np^{1/2})$ edges suffice to preserve p pairs [19] (which has the form $O(n + p)$ only at the trivial point $p = \Omega(n^2)$). However, on the negative side, it is not known how to improve *any* of the (undirected) lower bound constructions when edge directions are allowed. Thus, for example, techniques for reasoning about the directed setting are badly needed.

Another important setting studied by Coppersmith and Elkin is *subset distance preservers*:

Definition (Subset Distance Preservers – [19]). *When $P = S \times S$ for some node subset S in a graph G , a pairwise distance preserver of G, P is also called a subset distance preserver of G, S .*

Our picture of subset distance preservers is even fuzzier than our picture of pairwise distance preservers, in that our upper bounds do not benefit from the structure $P = S \times S$, but our lower bounds become much worse when this structure is required. Coppersmith and Elkin proved a host of lower bounds for subset preservers, which are summarized in Figure 2. Notably, our lower bounds on linear size subset preservers are *polynomially* far from $s = n$; that is, it is open whether or not we can always have $O(s^2)$ sized preservers when $s = \Omega(n^{9/16})$ (unweighted graphs) or $s = \Omega(n^{3/5})$ (weighted graphs). Confirming or refuting the existence of $O(s^2)$ size subset distance preservers when $s = n^{1-\varepsilon}$ for any absolute $\varepsilon > 0$ is perhaps the main open question in the field of distance preservers [19].

		Upper Bound	Lower Bound
Unwtd.	Undir.	Same as pairwise preservers (with s^2 substituted for p)	$\Omega(n^{10/11}s^{4/11} + n^{9/11}s^{6/11})$
	Dir.		$\omega(\mathbf{n} + \mathbf{s}^2)$ when $\omega(\mathbf{n}^{1/4}) = \mathbf{s} = \mathbf{o}(\mathbf{n}^{9/16})$
Wtd.	Undir.		$\Omega(n^{6/7}s^{4/7})$
	Dir.		$\omega(\mathbf{n} + \mathbf{s}^2)$ when $\omega(\mathbf{n}^{1/4}) = \mathbf{s} = \mathbf{o}(\mathbf{n}^{3/5})$

Figure 2: State of the art upper and lower bounds for subset distance preservers (of s nodes). All lower bounds are due to [19].

1.2 Our Results

We show three new existence/nonexistence results for linear size distance preservers, filling in several of the gaps mentioned above.

$O(n)$ -Sized Preservers for Directed Weighted Graphs. Our first result demonstrates the existence of nontrivial linear-sized preservers even in the most general setting of directed and weighted graphs. We show:

Theorem 1. *For any (possibly weighted and/or directed) graph G on n nodes, for any set P of p node pairs, there exists a distance preserver of G, P on $O(n + n^{2/3}p)$ edges.*

Thus, if $p = O(n^{1/3})$, then we have preservers of size $O(n)$. Previously, no nontrivial upper bounds were known even for directed *unweighted* graphs; i.e. it was conceivable that even any $\omega(1)$ pairs would require $\omega(n)$ edges for a distance preserver in the worst case. Our new upper bound improves on the previous upper bound of $O(np^{1/2})$ for directed weighted graphs whenever $p = o(n^{2/3})$.

This theorem is technically interesting due to its extremely short and simple proof. Our main new idea is a simple observation that extends a certain trick, based on counting “branchings” between paths, into the directed setting for the first time. Using this extension, we are able to re-apply a simple and elegant argument of [19], which was used to give upper bounds for distance preservers in undirected graphs, to obtain analogous bounds in the directed setting.

$O(p)$ -Sized Preservers for Undirected Unweighted Graphs. Our second result establishes the existence of preservers of size $O(p)$ when p is sufficiently large:

Theorem 2. *For any unweighted undirected graph G on n nodes, for any set P of p node pairs in G , if $p = \Omega\left(\frac{n^2}{\text{RS}(n)}\right)$ then there exists a preserver of G, P on $O(p)$ edges.*

Here, $\text{RS}(n)$ is the Ruzsa-Szemerédi function, which is defined such that any graph whose edges can be partitioned into n induced matchings has at most $\frac{n^2}{\text{RS}(n)}$ edges. The exact asymptotic value of $\text{RS}(n)$ is open, but it is known that

$$2^{\Omega(\log^* n)} = \text{RS}(n) = 2^{O(\sqrt{\log n \cdot \log \log n})}.$$

The lower bound is due to Fox [29] (improving on the original classic result by Ruzsa and Szemerédi [39]). The upper bound is implied via graphs built from dense sets that contain no arithmetic

progressions. The best such construction is due to Elkin [22] (improving on a classic construction of Behrend [9]). It would be a major breakthrough to substantially improve either of these bounds on $\text{RS}(n)$. Some of the notable constructions of Ruzsa-Szemerédi-like graphs include [4, 39, 28, 11]. As this is an old problem with a very rich history, it would be impractical to fully survey its importance in mathematics here; instead, we refer the reader to the survey by Conlon and Fox [18].

Theorem 2 gives the first nontrivial preservers on $O(p)$ edges known in any setting. Although our new preservers exist in a sub-polynomial range (no such theorem can hold in any polynomial range [19]), they immediately have a number of interesting consequences:

1. This theorem provides the first separation between any two settings for distance preservers, in particular between the weighted and unweighted settings. Recall that $\omega(p)$ edges are needed for undirected weighted graphs when $p = o(n^2)$. Here, we show that $O(p)$ edges suffice in some range with $p = o(n^2)$ for undirected *unweighted* graphs.
2. Recall that $\omega(n + p)$ edges are needed for preservers in undirected unweighted graphs, when $p = o\left(\frac{n^2}{2^{\Theta(\sqrt{\log n \cdot \log \log n})}}\right)$. Our new theorem implies that the range of this lower bound cannot be improved at all without also implying new upper bounds for the Ruzsa-Szemerédi function, which will likely be very difficult. This suggests that it may be beyond the scope of current mathematical techniques to improve the Coppersmith and Elkin distance preserver lower bounds in too serious a manner (if they are indeed improvable), since this may imply new upper bounds on $\text{RS}(n)$.
3. This theorem provides a new technical barrier for graph compression lower bounds. Recently, Abboud and Bodwin [2] showed that one cannot compress a graph into $o\left(\frac{n^{4/3}}{2^{\Theta(\sqrt{\log n \cdot \log \log n})}}\right)$ bits while preserving distance information up to any additive constant $+c$. The $n^{o(1)}$ factor in the denominator here is inherited from the fact that the construction uses the Coppersmith-Elkin lower bounds on $\omega(p)$ -size distance preservers as a central ingredient. In particular, recall that the Coppersmith-Elkin lower bounds on $\omega(p)$ -size preservers don't quite reach n^2 ; instead, they only go up to $n^{2-o(1)}$, missing by precisely the same $n^{o(1)}$ factor that appears in the Abboud-Bodwin compression lower bounds. It is tempting to try to improve the compression lower bound to $o(n^{4/3})$ by improving the $\omega(p)$ distance preserver lower bound up to $o(n^2)$. However, our new Theorem 2 shows that this is impossible; the $n^{o(1)}$ factor is inherent in distance preserver lower bounds and cannot be entirely removed. Thus, significant new ideas will be needed to improve the compression lower bound all the way to $o(n^{4/3})$, if true.
4. The proof of Theorem 2 overcomes a somewhat surprising technical hurdle. It was shown in [13] that, if one uses standard “consistent” methods for breaking ties between equally short paths, then no theorem of this sort will be possible; instead, the lower bound on $\omega(p)$ -size distance preservers goes all the way up to $o(n^2)$. In other words, any general construction of $O(p)$ size preservers when $p = o(n^2)$ (like the one offered by Theorem 2) must use some new unusual method of shortest path tiebreaking. In this paper we introduce “lazy tiebreaking,” which is used to sidestep this technical barrier. Lazy tiebreaking ensures that the resulting preserver will have a very particular shortest path structure that contains many induced matchings, which lets us draw our new connections between distance preservers and the Ruzsa-Szemerédi function.

Lower Bounds for Subset Distance Preservers Finally, we give new lower bounds for subset distance preservers:

Theorem 3. *There is a family of undirected weighted graphs G on n nodes and node subsets S of size $s = s(n)$ such that every subset distance preserver of G, S has $\Omega(sn^{2/3})$ edges.*

Theorem 4. *For any (possibly non-constant) integer $d \geq 2$, there is a family of undirected unweighted graphs G on n nodes and node subsets S of size $s = s(n)$ such that every subset distance preserver of G, S has*

$$\Omega \left(n^{2d/(d^2+1)} s^{(2d-1)(d-1)/(d^2+1)} \left(\frac{1}{2^{\Theta(\sqrt{\log n \cdot \log \log n})}} \right) \right)$$

edges.

These imply:

Corollary 1. *Subset distance preservers in (directed or undirected) weighted graphs need $\omega(s^2)$ edges in the worst case when $s = o(n^{2/3})$.*

Corollary 2. *Subset distance preservers (in directed or undirected and weighted or unweighted graphs) need $\omega(s^2)$ edges in the worst case when $s = o\left(\frac{n^{2/3}}{2^{\Theta(\sqrt{\log n \cdot \log \log n})}}\right)$.*

Both of these corollaries reflect a polynomial improvement in the lower bounds for s . In particular, the previous lower bounds ([19]) for weighted graphs stated that $\omega(s^2)$ edges were required when $s = o(n^{3/5})$, and for unweighted graphs only the range $s = o(n^{9/16})$ was established. More generally, Theorem 3 improves asymptotically over the previous lower bound for weighted subset preservers when $\omega(n^{4/9}) = s = o(n^{2/3})$, and Theorem 4 improves over the previous lower bound for unweighted subset preservers when $\omega(n^{1/3+o(1)}) = s = o(n^{2/3-o(1)})$.

Our proof technique is not similar to the one used for previous subset preserver lower bounds. While the bounds in [19] were proved using new graph constructions, our bounds are better viewed as a *reduction* from the subset to the pairwise setting. We adapt particular kind of graph replacement product used in [2] to compose instances of the Coppersmith and Elkin *pairwise* distance preserver lower bound graphs with themselves in a careful way. The resulting graph is a bit sparser than the original lower bound graphs, but its pair set P *nearly* has the structure $P = S \times S$. We can then take S to be the smallest node set that encloses P , and this only negligibly harms the quality of the final lower bound. Since the Coppersmith and Elkin pairwise preserver lower bound graphs are used as a direct ingredient in our proofs, in some sense, Corollaries 1 and 2 flow directly from the pairwise lower bounds discussed earlier, stating that $\omega(p)$ edges are needed in a pairwise preserver when $p = o(n^2)$ (weighted setting) or $p = o\left(n^2/2^{\Theta(\sqrt{\log n \cdot \log \log n})}\right)$ (unweighted setting).

1.3 Other Related Work

There has been lots of work on the relaxation of distance preservers in which distances must be preserved *up to an additive error term*. These subgraphs are called (*additive*) *pairwise spanners*, or sometimes *+k approximate distance preservers*. This line of work was implicit in many notable papers over the last 15 years, but it was first explicitly abstracted by Cygan, Grandoni, and Kavitha [20]. Later improved upper bounds came from Kavitha and Varma [31], Kavitha [30], and Bodwin and Williams [13], and new lower bounds came from Woodruff [41], Parter [32], and Abboud and Bodwin [1, 2].

More generally, the field of (*all pairs*) *spanners*, in which *all* pairwise distances must be approximately preserved, has received quite a lot of research attention over the last few decades. In this time, researchers have essentially fully understood what is possible in the regime of multiplicative

error [6, 7, 36, 27], nearly understood what is possible in the regime of additive error [3, 16, 8, 2], and provided some surprisingly strong upper bounds in the regime of mixed (i.e. both additive and multiplicative) error [23, 40, 37, 25, 8]. Another common variant is *fault-tolerant* spanners or distance preservers, which must (approximately) preserve distances even after some edges “fail.” Parter and Peleg [35] obtained matching upper and lower bounds for BFS structures in the presence of one fault, and Parter [34] obtained upper and lower bounds for the two fault case. Interesting fault-tolerant spanners were constructed in [17, 33, 21, 10, 15].

2 $O(n)$ -Sized Preservers for Directed Weighted Graphs

In this section, we show:

Theorem 1. *Any (possibly directed and/or weighted) graph $G = (V, E)$ and set P of p node pairs has a distance preserver on $O(n + n^{2/3}p)$ edges.*

Our first observation is that it suffices to show that any set of $p = O(n^{1/3})$ node pairs has a linear size distance preserver on $O(n)$ edges. Given an arbitrary set of node pairs, we may then divide the pair set into groups of size $O(n^{1/3})$, build a preserver of each group individually, and then take a union bound over the (linear) sizes of each individual preserver to obtain the claimed upper bound. We will thus focus here on this simpler problem. Our proof is morally similar to the proof in [19] that all undirected graphs have distance preservers on $O(n + n^{1/2}p)$ edges.

We begin with some useful background on *tiebreaking schemes* for shortest paths. When the pairs in P do not have unique shortest paths between them, it is necessary to break ties (i.e. choose one of the many shortest paths to include in the preserver) in some predefined way. This notion is formalized as:

Definition 1 (Shortest Path Tiebreaking Scheme – [13]). *Given a graph G , a shortest path tiebreaking scheme (or tiebreaking scheme for short) is a function π that maps ordered pairs of nodes (s, t) to a shortest path in G from s to t . When the underlying graph is clear from context, we will suppress the subscript G . For a set of node pairs P , we will commonly use $\pi(P)$ as a shorthand for $\bigcup_{p \in P} \pi(p)$.*

The following is a useful property of tiebreaking schemes:

Definition 2 (Consistency – [13]). *A tiebreaking scheme π is consistent if, for all nodes $w, x, y, z \in V$, if $x, y \in \pi(w, z)$ with $\text{dist}(w, x) < \text{dist}(w, y)$, then $\pi(x, y)$ is a subpath of $\pi(w, z)$.*

Note that all graphs admit a consistent tiebreaking scheme:

Claim 1. *For any graph G , there is a tiebreaking scheme π that is consistent in G .*

Proof. One such tiebreaking scheme can be obtained as follows: let Δ be the minimum positive difference between any two shortest path lengths, and then randomly perturb each edge weight in G by a number drawn uniformly at random from the interval $[0, \frac{\Delta}{3n}]$. Since each shortest path uses at most n edges, this process changes the length of any path by at most $\frac{\Delta}{3}$, and thus it cannot cause a previously non-shortest path between two nodes to become shortest. Additionally, with probability 1, no two sets of edges will have equal total weight; thus, shortest paths in the new graph G' with perturbed edge weights are unique. We now define π to be the tiebreaking scheme in G that chooses the (unique) shortest path between two nodes in the perturbed graph G' . \square

To construct distance preservers implementing Theorem 1, simply choose shortest paths between the pairs in P according to any consistent tiebreaking scheme. That is, our preserver is the graph $H = (V, E_H := \pi(P))$ for some consistent tiebreaking scheme π . We will now begin to prove upper bounds on the size of this edge set E_H . The proof is a counting argument, based on the following definition:

Definition 3 (Branching Triple). *A branching triple is a set of three distinct (directed) edges $\{e_1 = (u_1, v), e_2 = (u_2, v), e_3 = (u_3, v)\}$ that all enter the same node.*

Lemma 1. *H has at most $\binom{p}{3}$ branching triples.*

Proof. For each edge e in H , assign ownership of e to some pair $p \in P$ such that $e \in \pi(p)$. We will prove the claim by arguing that there do not exist any two branching triples

$$t = \{e_1 = (u_1, v), e_2 = (u_2, v), e_3 = (u_3, v)\} \text{ and } t' = \{e'_1 = (u'_1, v'), e'_2 = (u'_2, v'), e'_3 = (u'_3, v')\}$$

with their edges owned by the same set of three pairs p_1, p_2, p_3 (with e_i and e'_i owned by p_i for each $i \in \{1, 2, 3\}$). Suppose towards a contradiction that t, t' exist as described.

Assume without loss of generality that at least two of the three edges $e_i \in t$ precede the corresponding edge $e'_i \in t'$ in their respective paths p_i . More specifically, we assume that e_1 precedes e'_1 in p_1 and e_2 precedes e'_2 in p_2 . It follows that $v, v' \in \pi(p_1)$ (with v preceding v'), and also $v, v' \in \pi(p_2)$ (again with v preceding v'). Since π is consistent, this means that $\pi(v, v')$ is a subpath of both p_1 and p_2 . Therefore, p_1 and p_2 both contain the same edge entering v' , and so $e'_1 = e'_2$. However, by definition, all edges must be unique within any branching triple. By contradiction, then, t and t' cannot be owned by the exact same set of three pairs, and the lemma follows. \square

Claim 2. *A graph with $O(n)$ branching triples has $O(n)$ edges.*

Proof. Consider adding $O(n)$ edges one by one to an initially empty graph. The first and second edge entering a node do not create any new branching triples. Each subsequent edge creates (at least) one new branching triple. Therefore, the number of edges in the graph is at most $2n$ more than the number of branching triples in the graph. \square

It now follows straightforwardly that all G, P has a preserver on $O(n)$ edges whenever $|P| = O(n^{1/3})$, which implies Theorem 1.

3 $O(p)$ -Sized Preservers for Undirected Unweighted Graphs

The following definitions will be useful in this section:

Definition 4 (Induced Matching). *Given an undirected unweighted graph $G = (V, E)$, a set of edges $E' \subseteq E$ is an induced matching if E' is a matching, and there exists a set of nodes $S \subseteq V$ such that edges in the subgraph $G[S]$ of G induced on S are exactly E' .*

Definition 5 (Ruzsa-Szemerédi Graph). *A (undirected unweighted) graph $G = (V, E)$ is a Ruzsa-Szemerédi graph if its edge set can be partitioned into at most n induced matchings.*¹

¹Standard terminology is that a graph is an (r, t) -Ruzsa-Szemerédi graph if its edges can be partitioned into t induced matchings, each of size exactly r . In this paper we suppress the leading (r, t) because we exclusively consider $t = n$, and we omit the detail that the matchings must have the same size (which affects all relevant functions only by a constant factor).

Definition 6 ($\text{RS}(n)$). *The function $\text{RS}(n)$ is defined to be the largest value such that every Ruzsa-Szemerédi graph on n nodes has at most $\frac{n^2}{\text{RS}(n)}$ edges.*

See the introduction for more details on the function $\text{RS}(n)$. We will show:

Theorem 2. *Any undirected unweighted graph $G = (V, E)$ and set P of $p = \Omega\left(\frac{n^2}{\text{RS}(n)}\right)$ node pairs has a distance preserver on $O(p)$ edges.*

3.1 Reduction to the Bipartite Setting

For technical reasons, our main proof of the sparsity of H will require an assumption that the underlying graph is bipartite. Here, we show that this assumption can be made while essentially only paying negligible cost in the preserver density. A similar reduction, which contains different terminology and is not strong enough for our purposes here, appeared in [1] (although this reduction was used for a very different reason).

We remark that it is not strictly necessary to include this graph transformation as part of the preserver *construction*; that is, Theorem 2 applies even if we perform the construction in Section 3.2 to the original non-bipartite graph. We would then imagine performing the reduction in this section to the preserver solely for the sake of the analysis. However, it simplifies the proofs a bit (if not the construction) to preprocess the graph to be bipartite, so we present our construction in this order.²

Lemma 2. *Suppose that every distance preserver of a bipartite graph G and set P of p node pairs has at most $f(n, p)$ edges. Then every distance preserver of an arbitrary graph G and a set P of p node pairs has at most $f(2n, 2p)$ edges.*

This lemma holds in any setting; e.g. it is true both for directed weighted graphs and undirected unweighted graphs. In this paper, however, we will only apply it in the undirected unweighted setting.

Our main transformation is the *bipartite preserver lift*:

Definition 7 (Bipartite Preserver Lift). *Given a graph $G = (V, E)$ and a set of node pairs P , the bipartite preserver lift of G, P is a bipartite graph $G' = ((V_1, V_2), E')$ and a set of node pairs P' in G' defined as follows:*

- V_1 and V_2 are both identical copies of V . For any node $x \in V$, we will use the subscript x_1 (x_2) to denote the corresponding node in V_1 (V_2).
- For each edge $(u, v) \in E$, we include edges (u_1, v_2) and (u_2, v_1) in E' .
- For each pair $(s, t) \in P$, if $\text{dist}_G(s, t)$ is even, then we include pairs $(s_1, t_1), (s_2, t_2)$ in P' . If $\text{dist}_G(s, t)$ is odd, then instead we include pairs $(s_1, t_2), (s_2, t_1)$ in P' .

We call the (nearly) inverse operation to the bipartite lift a *contraction*:

Definition 8 (Contraction). *Let $G' = (V_1 \cup V_2, E'), P'$ be the bipartite preserver lift of some $G = (V, E), P$, and let $H' = (V_1 \cup V_2, E'_H) \subseteq G'$. The contraction of H' is a graph $H = (V, E_H) \subseteq G$, where*

$$E_H := \{(u, v) \in V \times V \mid (u_1, v_2) \in E'_H \text{ or } (u_2, v_1) \in E'_H\}.$$

²This observation was suggested by an anonymous reviewer.

Claim 3. *Let G', P' be the bipartite lift of some G, P . Then for any pair $(s_i, t_j) \in P'$ (where $i, j \in \{1, 2\}$), we have $\text{dist}_{G'}(s_i, t_j) \leq \text{dist}_G(s, t)$.³*

Proof. We will prove the claim for the case $i = j = 1$; the other settings of i, j follow from a symmetric argument.

Let $q = (s, x^1, \dots, x^{k-1}, t)$ be a shortest path from s to t in G . Note that by construction, the existence of $(s_1, t_1) \in P'$ implies that $\text{dist}_G(s, t)$ is even, and so $k - 1$ is odd. Therefore, we have a path mirroring q of the form $(s_1, x_2^1, \dots, x_2^{k-1}, t_1)$ in G' , and so there exists a path in G' of length k from s_1 to t_1 . Thus, $\text{dist}_{G'}(s_1, t_1) = k \geq \text{dist}_G(s, t)$. \square

We now have:

Claim 4. *Let $G = (V, E)$ be a graph and let P be a set of node pairs in G . Then if H' is a preserver of the bipartite preserver lift G', P' of G, P , then the contraction H of H' is a preserver of G, P .*

Proof. Consider any pair $(s, t) \in P$. Assume that $\text{dist}_G(s, t)$ is even; the case in which $\text{dist}_G(s, t)$ is odd follows from a symmetric argument. By construction we have $(s_1, t_1) \in P'$, and by Claim 3 we know that $\text{dist}_{G'}(s_1, t_1) \leq \text{dist}_G(s, t)$. Thus, there exists a path from s_1 to t_1 in H' of length at most $\text{dist}_G(s, t)$. Let

$$q' := (s_1, x_2^1, \dots, x_2^{k-1}, t_1)$$

be this path. By construction, there exists a path

$$q = (s, x^1, \dots, x^{k-1}, t)$$

in H . It follows that $\text{dist}_H(s, t) \leq \text{dist}_G(s, t)$. To complete the proof, we will argue that $H \subseteq G$, and so $\text{dist}_H(s, t) = \text{dist}_G(s, t)$ and H is a preserver of G, P . The fact that $H \subseteq G$ follows quite simply from the definitions of preserver lifts and contractions: if there is an edge (u, v) in H , then there is an edge (u_1, v_2) or (u_2, v_1) in H' ; since $H' \subseteq G'$ this edge is also in G' , and thus we have that (u, v) is an edge in G . \square

Given our preserver H , we then arbitrary graph G , we may then prove Lemma 2 by computing a bipartite lift G', P' of G, P , building a preserver H' of G', P' on $f(2n, 2p)$ edges, and then contracting H' back down to a graph H on at most $f(2n, 2p)$ edges which is a preserver of G, P .

We proceed with the assumption that the underlying graph G in Theorem 2 is bipartite.

3.2 Lazy Tiebreaking

We will first describe the tiebreaking scheme used to build our preserver. First, we need some new notation.

Definition 9 (P_s). *Given a graph $G = (V, E)$ and a set of node pairs P , for any node $s \in V$, the set P_s is defined as the subset of pairs in P with source node s . That is, $P_s := \{(s, t) \in P \mid t \in V\}$.*

Definition 10 (Branching Edges). *In a tree T rooted at a node s , a node b is a branching node if, when the edges of T are oriented away from s , we have $\deg_{\text{out}}(b) \geq 2$. Any edge leaving a branching node b under this orientation of T is then called a branching edge. The set of branching edges of T is denoted $\mathcal{B}(T)$.*

Definition 11 (Lazy Tiebreaking). *A tiebreaking scheme π in a graph $G = (V, E)$ is called lazy for a pair set P if:*

³In fact $\text{dist}_{G'}(s_i, t_j) = \text{dist}_G(s, t)$, but we will only need to show this weaker fact.

1. For all $s \in V$, the graph $T_s := (V, \pi(P_s))$ is a tree (possibly together with some isolated nodes), and
2. For all distinct non-branching edges $(x, y), (x', y') \in T_s \setminus \mathcal{B}(T_s)$ with $\text{dist}_G(s, y) = \text{dist}_G(s, y') = \text{dist}_G(s, x) + 1 = \text{dist}_G(s, x') + 1$, we have $(x, y') \notin E$ (and, by symmetry, $(x', y) \notin E$).

Informally, a tiebreaking scheme is “lazy” if it tries to delay the branching edges of T_s for as long as possible. This follows from the definition because, if the second requirement of lazy tiebreaking is violated – i.e. $(x', y) \in E$ – then we can re-choose the path from s to y such that it passes through x instead of x' , and this will delay its final branching edge.

All graphs and sets of node pairs admit a lazy tiebreaking scheme:

Claim 5. *For any graph G and set of node pairs P , there is a tiebreaking scheme π in G that is lazy for P .*

Proof. Note from the definition of lazy tiebreaking that it suffices to only consider a pair set $P = P_s$ in which all pairs have the same source node s . Let \mathcal{T} be the set of all preservers of P_s in G that are trees (possibly together with some isolated nodes). For any $T_s \in \mathcal{T}$, sort the branching edges $\mathcal{B}(T_s)$ in descending order by distance from s , and let $\mathcal{D}(T_s)$ be the (descending) list of these distances. We now define a partial ordering over the elements of \mathcal{T} : say that $T_s < T'_s$ if $\mathcal{D}_i(T_s) < \mathcal{D}_i(T'_s)$ for the first index i on which the two lists differ (if the lists are identical, or one is a prefix of the other, then T_s, T'_s are incomparable). Now choose any maximal element T_s^{\max} in this partial ordering. We will argue that the corresponding tiebreaking scheme π such that $\pi(P_s) = T_s^{\max}$ is lazy.

To see this, assume towards a contradiction that π is *not* lazy; that is, there are distinct non-branching edges $(x, y), (x', y') \in T_s^{\max} \setminus \mathcal{B}(T_s^{\max})$ with $\text{dist}_G(s, y) = \text{dist}_G(s, y') = \text{dist}_G(s, x) + 1 = \text{dist}_G(s, x') + 1$ and $(x, y') \in E$. We may then define a new tiebreaking scheme π' , which reroutes all paths containing the edge (x', y') such that they now follow T_s^{\max} from s to x , then walk the edge (x, y') , and then continue along their old suffix to their non- s endpoint (for pairs that do not use the edge (x', y') under π , we define π' to agree with π). Note that in $\pi'(P_s)$, the edge (x, y) is a branching edge, whereas it is not a branching edge in $\pi(P_s)$. Further, $\pi(P_s)$ and $\pi'(P_s)$ have exactly the same set of edges at distance further than (x, y) from s . Additionally, observe that (x', y') is the only edge that is equally far from s as (x, y) that belongs to $\pi(P_s)$ but not $\pi'(P_s)$. Since we have assumed that (x', y') is not a branching edge, these three conditions imply that $\pi'(P_s) > \pi(P_s) = T_s^{\max}$. This contradicts the assumption that T_s^{\max} is a maximal element in the partial ordering; thus, π is lazy. \square

A preserver implementing Theorem 2 may be built using any lazy tiebreaking scheme: that is, our preserver is $H := (V, E_H := \pi(P))$ for some π that is lazy for P .

3.3 Proof of Correctness

Lemma 3. *The edges of $E_H \setminus \bigcup_{s \in V} \mathcal{B}(T_s)$ can be partitioned into $3n$ induced matchings.*

Proof. For each edge

$$e = (u_1, v_2) \in E_H \setminus \bigcup_{s \in V} \mathcal{B}(T_s)$$

assign e to a node s such that $e \in T_s$. Next, for each $s \in V$, partition the edges owned by s into three equivalence classes C_s^0, C_s^1, C_s^2 . Each edge (u, v) , with $\text{dist}_G(s, u) < \text{dist}_G(s, v)$, is assigned to the class C_s^i where $i := \text{dist}_G(s, u) \bmod 3$. There are $3n$ such classes in total. We will now show that each class is an induced matching.

Let $(u, v), (w, x) \in C_s^i$ for some $s \in V, i \in \{0, 1, 2\}$. Assume without loss of generality that $\text{dist}_G(s, u) \leq \text{dist}_G(s, w)$. We split into two cases:

First, suppose $\text{dist}_G(s, u) + 3 \leq \text{dist}_G(s, w)$, and so we additionally have $\text{dist}_G(s, u) + 4 \leq \text{dist}_G(s, x)$. Then none of the edges $(u, w), (u, x), (v, w), (v, x)$ may exist in G , as any one of these edges would imply the existence of a path from u to x of length 3 or less.

Since $(u, v), (w, x) \in C_s^i$ for some i , the only other case to consider is when $\text{dist}_G(s, u) = \text{dist}_G(s, w)$, and thus also $\text{dist}_G(s, v) = \text{dist}_G(s, x)$. In this case, we first observe that $(u, w), (v, x) \notin E$, since this would give us an odd cycle $s \rightsquigarrow u \rightarrow w \rightsquigarrow s$ or $s \rightsquigarrow v \rightarrow x \rightsquigarrow s$ (which cannot exist since G is bipartite). Second, we note that $(u, x), (v, w) \notin E$ by the second property of lazy tiebreaking.

We now have that $(u, w), (u, x), (v, w), (v, x) \notin E$. Consider the graph induced on the endpoints of all edges in some particular C_s^i . It follows that the edge set of this graph is *precisely* the set of edges in C_s^i , and so each C_s^i is an induced matching. \square

Claim 6. $\left| \bigcup_{s \in V} \mathcal{B}(T_s) \right| = O(p)$

Proof. For any fixed s , the tree T_s has at most $|P_s|$ leaves; thus, by standard structure of trees, it has $O(|P_s|)$ branching edges. We then have

$$\left| \bigcup_{s \in V} \mathcal{B}(T_s) \right| = \sum_{s \in V} O(|P_s|) = O(p)$$

where the second equality follows from the fact that the collection $\{P_s\}$ over all $s \in V$ is a partition of P .⁴ \square

Lemma 4.

$$|E_H| = O(p) + O\left(\frac{n^2}{\text{RS}(n)}\right)$$

Proof. Define a new graph $R \subseteq H$ as follows. First, remove all edges in $\bigcup_{s \in V} \mathcal{B}(T_s)$ from E_H . By Lemma 3, the remaining graph can be partitioned into $3n$ induced matchings; keep all edges in the n largest matchings (breaking ties arbitrarily) and discard the rest. Clearly we discard only a constant fraction of the edges in this way. The remaining edge set E_R can be partitioned into n induced matchings, so it is a Ruzsa-Szemerédi graph. Thus, we have

$$|E_R| \leq \frac{n^2}{\text{RS}(n)}$$

and so

$$|E_H| = O\left(\frac{n^2}{\text{RS}(n)}\right) + \left| \bigcup_{s \in V} \mathcal{B}(T_s) \right|.$$

By Claim 6, we then have

$$|E_H| = O\left(\frac{n^2}{\text{RS}(n)}\right) + O(p).$$

\square

⁴It is convenient here to interpret the pairs in P as ordered, so that a single pair $(s, t) \in P$ is not considered in both P_s and P_t .

We can now show:

Proof of Theorem 2. Let $f(n, p)$ be the maximum number of edges needed in a preserver of an undirected unweighted bipartite graph on n nodes and a set of p node pairs. By Lemma 4, we have

$$f(n, p) = O\left(\frac{n^2}{\text{RS}(n)}\right) + O(p)$$

By Lemma 2, the maximum number of edges needed in a preserver of a (not necessarily bipartite) graph on n nodes and a set of p node pairs is

$$f(2n, 2p) = O\left(\frac{(2n)^2}{\text{RS}(2n)}\right) + O(2p) = O\left(\frac{n^2}{\text{RS}(2n)}\right) + O(p)$$

Clearly $\text{RS}(2n) \geq \text{RS}(n)$, and so we have

$$f(2n, 2p) = O\left(\frac{n^2}{\text{RS}(n)}\right) + O(p)$$

Thus, if $p = \Omega\left(\frac{n^2}{\text{RS}(n)}\right)$, then we have $f(2n, 2p) = O(p)$, which completes the proof. \square

4 Lower Bounds for Subset Preservers

We show:

Theorem 3. *For any $s = s(n)$, there is a family of undirected weighted graphs $G = (V, E)$ and sets S of s nodes in G such that any subset distance preserver of G, S has $\Omega(n^{2/3}s)$ edges.*

Theorem 4. *For any integer $d \geq 2$, for any $s = s(n)$, there is a family of undirected unweighted graphs $G = (V, E)$ and sets S of s nodes in G such that any subset distance preserver of G, S has*

$$\Omega\left(n^{2d/(d^2+1)} s^{(2d-1)(d-1)/(d^2+1)} \left(\frac{1}{2^{\Theta(\sqrt{\log n \cdot \log \log n})}}\right)\right)$$

edges.

These imply:

Corollary 1. *Subset distance preservers in (directed or undirected) weighted graphs need $\omega(s^2)$ edges in the worst case when $s = o(n^{2/3})$.*

Proof. Immediate from Theorem 3. \square

Corollary 2. *Subset distance preservers (in directed or undirected and weighted or unweighted graphs) need $\omega(s^2)$ edges in the worst case when*

$$s = o\left(\frac{n^{2/3}}{2^{\Theta(\sqrt{\log n \cdot \log \log n})}}\right).$$

Proof. By direct calculation, the lower bound in Theorem 4 is $\omega(s^2)$ when

$$s = \omega \left(n^{2d/(3d+1)} / 2^{\Theta(\sqrt{\log n \cdot \log \log n})} \right) = \omega \left(n^{2/3-2/(9d+3)} / 2^{\Theta(\sqrt{\log n \cdot \log \log n})} \right).$$

We now choose $d = \Theta \left(\sqrt{\log n / (\log \log n)} \right)$, and so

$$n^{2/(9d+3)} = n^{1/\Theta(\sqrt{\log n / (\log \log n)})} = n^{\Theta(\sqrt{(\log \log n) / \log n})} = 2^{\Theta(\sqrt{\log n \cdot \log \log n})}$$

Thus, the lower bound is $\omega(s^2)$ when

$$s = \omega \left(n^{2/3} / 2^{\Theta(\sqrt{\log n \cdot \log \log n})} \right)$$

and the corollary follows. \square

We now give the proofs of Theorems 3 and 4. We begin in the weighted setting, which is a bit simpler.

4.1 The Weighted Construction

We begin the discussion by recalling the weighted *pairwise* lower bounds of Coppersmith and Elkin:

Theorem 5 ([19]). *For any n , there is a graph G on n nodes and a set P of $p = p(n)$ node pairs such that any distance preserver of G, P has $\Omega(n^{2/3}p^{2/3})$ edges.*

Let us first give an informal overview of the proof. Our goal is to convert the Coppersmith and Elkin pairwise lower bound graph into a lower bound for subset preservers. It is tempting to simply define S to be the set of nodes consisting of all endpoints in P ; then clearly a subset preserver of G, S will need $\Omega(n^{2/3}p^{2/3})$ edges. Unfortunately, a closer look at the Coppersmith and Elkin construction reveals that this straightforward approach will fail:

- We suffer some “cost” for using $S \times S$ in place of P , since all pairs in $(S \times S) \setminus P$ are “useless” – i.e. we must count them in our pair set, but they are not being used to enforce that any particular shortest path remains in the lower bound graph. Let $L = \frac{s^2}{p}$ capture this cost.
- The “value” of a useful pair in P is the number of edges that it adds to the preserver lower bound. It turns out that the value of each pair in P is L (up to constant factors). In other words, the cost *exactly* offsets the value in this approach; that is, the *average* value of a pair in $S \times S$ is only $O(1)$, so this naive construction only yields the trivial lower bound of $\Omega(s^2)$.

We fix the construction by applying the *obstacle product* technique of [2] (see Figure 3). The main idea here is that we can replace a single node v in a distance preserver lower bound graph with a copy G_v of a distance preserver lower bound graph, and the result is still a valid distance preserver lower bound graph. In our construction, we use the obstacle product to replace only *intermediate* nodes (i.e. nodes *not* in S) in the Coppersmith and Elkin pairwise lower bound graph. Thus, the “cost” of applying the structure $P = S \times S$ remains unchanged, but our useful pairs must now pass through $\approx L$ copies of G_v rather than $\approx L$ nodes. Since it takes a super-constant number of edges to cross a copy of G_v , our value improves enough that we obtain an interesting subset preserver lower bounds.

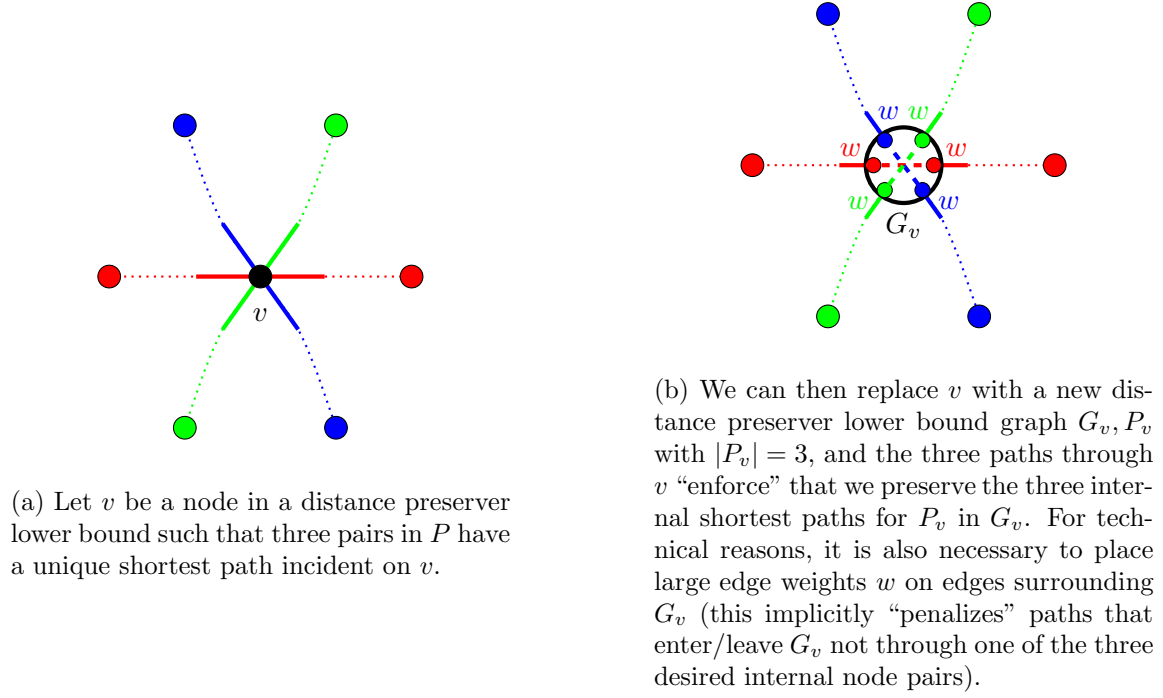


Figure 3: The Obstacle Product technique lets us replace nodes in a distance preserver lower bound graph with entire new lower bound graphs.

We now give the construction in more formality. We first define an *inner graph* and an *outer graph*, where the outer graph is the original pairwise preserver lower bound, and the inner graph will be used to replace some of the nodes of the outer graph. The outer graph is always the following instantiation of Theorem 5 (regardless of the desired size of s):

Theorem 6 ([19]). *There is an infinite family of 3-layered undirected weighted graphs G with n nodes per layer and $\Omega(n^2)$ edges, as well as a set P_O of p_O node pairs with one point in the first layer and the other point in the last layer, such that:*

1. *All nodes in the middle layer have the same even degree $D = \Theta(n)$.*
2. *For all $(s, t) \in P_O$, there is a unique shortest path in G from s to t , and it includes exactly two edges.*
3. *For each edge in G , there is a unique pair in P_O whose shortest path includes that edge.*

Our outer graph is precisely one of these three-layered graphs. We now modify our outer graph by replacing every node in the middle layer with an *inner graph* drawn from Theorem 5. The specific process of replacing a node with an inner graph proceeds as follows. Let v be a node in the middle layer of the outer graph which we desire to replace. Let the inner graph G_I^v be a distance preserver lower bound graph from Theorem 5, instantiated with:

- Pair set P_I^v of size $|P_I^v| = p_I = \frac{D}{2}$ node pairs and
- I nodes, for some $I \geq \Omega(D^2)$ that is a parameter of the construction.

We remove v from the graph and insert a copy of G_I called G_I^v into the graph. For each edge (c, v) that used to be incident on v , we replace the edge with (c, v') for some node $v' \in G_I^v$, chosen by the following process:

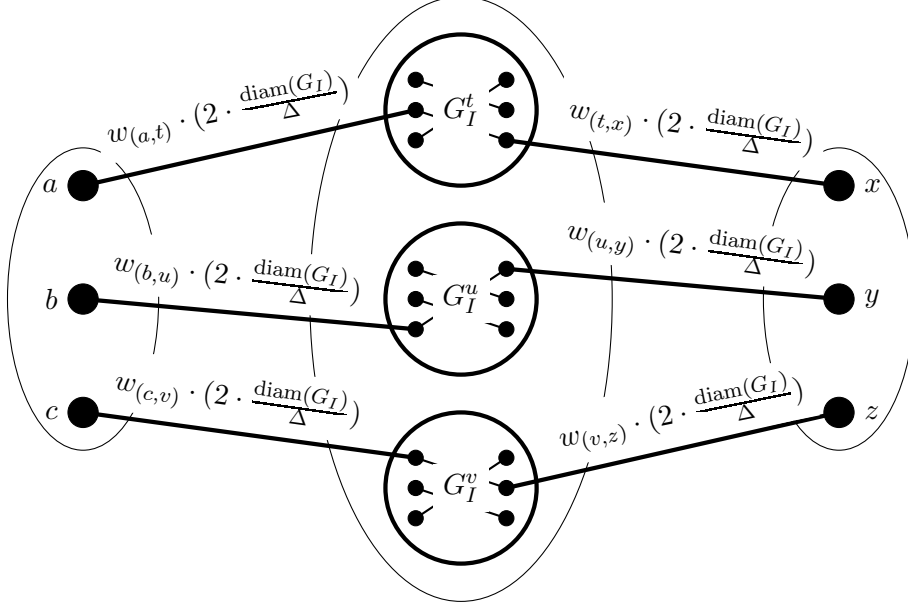


Figure 4: The construction of our subset distance preserver lower bound. Only some of the paths from the first layer of the outer graph to the last layer of the outer graph are pictured here.

By Theorem 6, we may partition the D edges incident on v into $\frac{D}{2}$ pairs, where for each pair of edges $\{(c, v), (v, z)\}$ we have $(c, z) \in P_O$ and the unique shortest $c \rightsquigarrow z$ path consists of exactly these two edges. For each such pair of edges $\{(c, v), (v, z)\}$, choose any distinct node pair $(q, r) \in P_I^v$, and replace the old edges $(c, v), (v, z)$ with new edges $(c, q), (r, z)$. The weights of these edges remain unchanged (for now), i.e. $w_{(c,v)} = w_{(c,q)}$ and $w_{(v,z)} = w_{(r,z)}$. We replace every node v in the middle layer of the outer graph with a distinct copy G_I^v in this manner.

One step remains in the construction. Intuitively, we hope that the pair $(c, z) \in P_O$ still has a unique shortest path, and that this shortest path enters/leaves G_I^v by some pair $(q, r) \in P_I^v$, thus enforcing that the shortest $q \rightsquigarrow r$ path in G_I^v must stay in the preserver. Unfortunately, this might not be the case in the construction so far: it could be that a new shortest $c \rightsquigarrow z$ path veers off course and enters/leaves G_I^v in a different way.

To solve this problem, we simply observe that any such path must correspond to a suboptimal route from c to z in the outer graph. Thus, if we simply scale up the edge weights of the outer graph to be much larger than the edge weights of the inner graph, then there is a huge implicit penalty for taking this suboptimal path. With large enough weights, the penalty becomes so high that we can show that no such alternate path is possible, and the problem is solved. We will see in calculations that follow in the next subsection that the following scaling is sufficient: let Δ be the minimum positive difference between the lengths of any two shortest paths in G_O (before any replacement occurs); we then multiply the weight of all *outer* edges – i.e. edges that are not contained in any copy G_I^v – by a factor of $2 \cdot \frac{\text{diam}(G_I)}{\Delta}$. Note that scaling by any larger factor would work just as well.

We will call the final graph at the end of the construction G . The node subset S whose distances must be preserved is defined to be all nodes in the first layer of G_O together with all nodes in the last layer of G_O . We then have $P_O \subseteq S \times S$, so we may simply argue that this graph cannot be sparsified below a certain level without increasing a pairwise distance in P_O .

4.2 Proof of Correctness

Our graphs have $nI + 2n$ nodes, since there are $2n$ nodes in the first and last layer of G_O , and each of the n nodes in the middle layer has been replaced with a subgraph on exactly I nodes. Our node subset S then has size $2n$. We will next argue that any subset distance preserver H of G, S must keep at least $\Omega(I^{2/3}n^{2/3})$ edges in *each* copy of G_I , which totals $\Omega(I^{2/3}n^{5/3})$ edges in H . Note that, if $N = \Theta(nI)$ is the number of nodes in G , and $s = \Theta(n)$ is the number of nodes in S , then this means that H has $\Omega(N^{2/3}s)$ edges, which implies Theorem 3.

First, we observe that the shortest paths for P_O in the final construction G have the desired structure.

Lemma 5. *Let $(c, z) \in P_O$, and let $\{(c, v), (v, z)\}$ be the unique shortest path from c to z in G_O . There exists a node pair $(q, r) \in P_I^v$ such that every shortest path from c to z in G walks the edge (c, q) , then walks a shortest path from q to r in G_I^v , and finally walks the edge (r, z) .*

Proof. First, by construction, there certainly *exists* a $c \rightsquigarrow v$ path Q in G with the structure described in this lemma statement. The length of Q is at most

$$|Q| \leq 2 \cdot \frac{\text{diam}(G_I)}{\Delta} (w_{(c,v)} + w_{(v,z)}) + \text{diam}(G_I).$$

Now consider any path Q' from c to z in G that does *not* have the form described in this lemma. If Q' walks the edge (c, q) , then walks a *non*-shortest path in G_I^v from q to r in G_I^v , then walks the edge (r, z) , it is obvious that $|Q'| > |Q|$ and so Q' is a non-shortest path from c to z in G . The more interesting case to consider is when Q' has the following form: it starts at c , then walks through some inner graph $G_I^{x_1}$, then it walks to a node y_1 (which could be in the first or last layer of G), then it walks through some inner graph $G_I^{x_2}$, and so on until it finally walks through an inner graph $G_I^{x_k}$ and then walks to its destination z . We may *underestimate* the length of this path, ignoring the edges used by this path in any inner graph, by the following expression:

$$|Q'| \geq 2 \cdot \frac{\text{diam}(G_I)}{\Delta} \left(w_{(c,x_1)} + \sum_{i=1}^{k-1} (w_{(x_i,y_i)} + w_{(y_i,x_{i+1})}) + w_{(x_k,z)} \right)$$

Note that the sub-expression

$$w_{(c,x_1)} + \sum_{i=1}^{k-1} (w_{(x_i,y_i)} + w_{(y_i,x_{i+1})}) + w_{(x_k,z)}$$

describes the length of some non-shortest path from c to z in G_O . Therefore, by definition of Δ , the length of this path is at least Δ more than $\text{dist}_{G_O}(c, z) = w_{(c,v)} + w_{(v,z)}$. We thus have:

$$\begin{aligned} |Q'| &\geq 2 \cdot \frac{\text{diam}(G_I)}{\Delta} (w_{(c,v)} + w_{(v,z)} + \Delta) \\ &= 2 \cdot \frac{\text{diam}(G_I)}{\Delta} (w_{(c,v)} + w_{(v,z)}) + 2 \cdot \text{diam}(G_I) \\ &> 2 \cdot \frac{\text{diam}(G_I)}{\Delta} (w_{(c,v)} + w_{(v,z)}) + \text{diam}(G_I) \\ &\geq |Q| \end{aligned}$$

So Q is strictly shorter than Q' , and Q' is not a shortest path in G . The lemma follows. □

Claim 7. *For any node pair $(q, r) \in P_I^v$ (for any v), there is a pair $(c, z) \in P_O$ such that every shortest path from c to z in G includes as a subpath some shortest path from q to r in G_I^v .*

Proof. Let $P_O^v \subseteq P_O$ be the set of node pairs $(c, z) \in P_O$ whose unique shortest path in G_O includes the node v . It follows from Theorem 6 that there are exactly $\frac{D}{2}$ such pairs in P_O^v . When we replace v with the inner graph G_I^v , we also have exactly $p_I = \frac{D}{2}$ pairs in P_I^v . Thus, we create a one-to-one correspondence between the pairs in P_O^v and the pairs in P_I^v . In other words, for any given pair $(q, r) \in P_I^v$, there is a pair $(c, z) \in P_O^v$ such that the edges of the shortest path $\{(c, v), (v, z)\}$ in G_O are replaced by edges $(c, q), (r, z)$ in G . We may now apply Lemma 5 to conclude that every shortest path from c to z in G includes as a subpath some shortest path from q to r in G_I^v . \square

Claim 8. *Any distance preserver H of G, P_O must keep $\Omega(I^{2/3}n^{2/3})$ edges in each inner graph G_I^v .*

Proof. Suppose towards a contradiction that H has $o(I^{2/3}n^{2/3})$ edges in some inner graph G_I^v . From Theorem 5, there exists a node pair $(q, r) \in P_I^v$ for which *no* shortest path from q to r in G_I^v has all of its edges still in H . By Claim 7, there is a pair $(c, z) \in P_O$ for which every shortest path from c to z in G includes as a subpath some shortest path from q to r in G_I^v . It follows that no shortest path from c to z in G has all of its edges remaining in H . This contradicts the fact that H is a preserver of G, P_O , and thus, H must keep $\Omega(I^{2/3}n^{2/3})$ edges in each inner graph G_I^v . \square

Theorem 3 now follows from the calculations at the beginning of this subsection.

4.3 The Unweighted Construction

The intuition behind the construction in the unweighted setting is exactly the same as the intuition in the weighted setting. However, a problem arises due to our inability to use edge weights: while previously we placed large edge weights on our “outer edges” to enforce that the shortest path topology of the graph endured over the obstacle product replacement, no such operation is possible in the unweighted setting (the natural attempt is to replace weighted edges with long paths, but this introduces too many new nodes to the graph and critically harms the lower bound). Instead, we must use some caution in our construction to avoid changing the shortest paths of the graph over the inner graph replacement step.

The solution we employ is to simply use a layered inner graph for which all pairs $(q, r) \in P_I^v$ have q in the first layer and r in the last layer. We then perform the obstacle product replacement as before, but we take care to only connect first-layer inner graph nodes to first-layer outer graph nodes, and last-layer inner graph nodes to last-layer outer graph nodes. Intuitively, this suffices because *any* $c \rightsquigarrow z$ path for a pair $(c, z) \in G_O$ must then cross a system of inner graph layers *at some point* just to get from the first to the third layer of the outer graph. Thus, we can argue that no alternate $c \rightsquigarrow z$ path can be shorter than the one that simply takes a direct route across the layers of G_I^v .

We now proceed with more formality. We draw our outer graphs from the following theorem:

Theorem 7 ([19]). *There is an infinite family of 3-layered undirected unweighted graphs G with n nodes per layer and*

$$\Omega\left(\frac{n^2}{2^{\Theta(\sqrt{\log n \cdot \log \log n})}}\right)$$

edges, as well as a set P_O of p_O node pairs with one point in the first layer and the other point in the last layer, such that:

1. *All nodes in the middle layer have the same even degree $D = \Theta\left(\frac{n}{2^{\Theta(\sqrt{\log n \cdot \log \log n})}}\right)$.*

2. For all $(s, t) \in P_O$, there is a unique shortest path in G from s to t , and it includes exactly two edges.
3. For each edge in G , there is a unique pair in P_O whose shortest path includes that edge.

The following theorem is the analogue of Theorem 5:

Theorem 8 ([19]). *For any n , for any integer $d \geq 2$, there is an undirected unweighted graph G on n nodes and*

$$\Theta\left(n^{2d/(d^2+1)} p^{d(d-1)/(d^2+1)}\right)$$

edges, and a set P of $p = p(n)$ node pairs such that there is a unique shortest path in G between any pair in P , these shortest paths are edge disjoint, and the edges of G are precisely the union of these paths.

However, we do *not* draw our inner graphs directly from this theorem, since these graphs are not layered. Instead, we use:

Lemma 6. *For any n , for any integer $d \geq 2$, there is an undirected unweighted graph G on n nodes and*

$$\Theta\left(n^{2d/(d+1)^2} p^{d/(d+1)}\right)$$

*edges, and a set P of $p = p(n)$ node pairs such that there is a unique shortest path in G between any pair in P , these shortest paths are edge disjoint, and the edges of G are precisely the union of these paths. **Moreover**, there is an integer $\ell = \ell(n)$ such that G is a graph on ℓ layers and for each pair $(q, r) \in P$, we have q in the first layer, r in the last layer, and $\text{dist}_G(q, r) = \ell - 1$.*

Proof. Start with a graph G and pair set P from Theorem 8 with n nodes and p pairs. We modify G, P in two steps to produce a graph satisfying this lemma.

Regularizing P . First, we will slightly modify the pair set P such that all pairs have exactly the same distance in G . Let L be the average distance between a pair in P . Since our pairs have unique edge-disjoint shortest paths, note that there are exactly pL edges in G . For each pair $(q, r) \in P$, we remove (q, r) from P , we divide the shortest path from q to r into subpaths of length exactly $\lfloor L/2 \rfloor$, and we add the endpoints of each of these subpaths back to P as a new pair. If there is a shorter “remainder” subpath of length less than $\lfloor L/2 \rfloor$ left over at the end, we remove all edges in this subpath from G (and we do *not* add its endpoints back to P). It is clear that all pairs in P now have the same distance $\lfloor L/2 \rfloor$ between them. Further, the shortest paths in the new P are still unique and edge-disjoint, and G is precisely the union of these paths. We have removed at most $p \cdot (\lfloor L/2 \rfloor - 1)$ edges from G ; since G originally had $p \cdot L$ edges, a constant fraction of the edges in G remain. Finally, since each pair in P has distance exactly $\lfloor L/2 \rfloor$, and since each pair in P has a unique edge-disjoint shortest path in G between its endpoints, the total number of pairs in the new P must still be $\Theta(p)$. Thus, we have changed the size of P and the density of G only by constant factors throughout this process.

Layering G . Next, we make $\ell := \lfloor L/2 \rfloor + 1$ copies of G , and for each edge (u, v) in the original graph, we add an edge between the copy of u in layer i and the copy of v in layer $i + 1$ (for all $1 \leq i \leq \lfloor L/2 \rfloor$; we also add edges from the copy of v in layer i to the copy of u in layer $i + 1$). We replace each pair $(q, r) \in P$ with the pair $(q_1, r_{\lfloor L/2 \rfloor + 1})$, where q_1 is the copy of q in the first layer and $r_{\lfloor L/2 \rfloor + 1}$ is the copy of r in the last layer. Note that the total number of pairs does not change in this process. Also note that, by construction, since the original shortest paths were unique and edge-disjoint, the shortest paths for the pairs $(q_1, r_{\lfloor L/2 \rfloor + 1})$ are still unique and edge-disjoint. Additionally, these paths have length exactly $\lfloor L/2 \rfloor = \ell - 1$, as claimed.

Analysis. The graph G is now a layered graph on $\Theta(nL)$ nodes, as well as $\Theta(p)$ node pairs, and for each node pair there is a unique edge disjoint shortest path of length $\Theta(L)$ between its endpoints. Thus, any preserver of G, P must have $\Theta(pL)$ edges in it. Recall that pL is precisely equal to the number of edges originally in G . We may thus compute:

$$L = \frac{|E(G)|}{p} = \Theta(n^{2d/(d^2+1)} p^{d(d-1)/(d^2+1)-1}) = \Theta(n^{2d/(d^2+1)} p^{(-d-1)/(d^2+1)})$$

The number of edges in any preserver of G, P is then

$$\Theta(pL) = \Theta\left(n^{2d/(d^2+1)} p^{d(d-1)/(d^2+1)}\right) = \Theta\left((nL)^{2d/(d+1)^2} p^{d/(d+1)}\right)$$

and the lemma is complete. \square

We draw the inner graphs in our construction from Lemma 6. We then perform our replacement product as before, with the following essential detail discussed earlier: Given a shortest path $\{(c, v), (v, z)\}$ in G_O , when we replace v with G_I^v and choose a corresponding pair $(q, r) \in P_I^v$, we add edges (c, q) and (r, z) to our new graph *where c, q are in the first layers of G_O, G_I^v (respectively) and z, r are in the last layers of G_O, G_I^v (respectively).*

The fact that we use layered inner graphs in this way allows us to re-prove Lemma 5 using a slightly modified argument:

Proof of Lemma 5 in the unweighted construction. As in the weighted construction, for some pair $(c, z) \in P_O$, let Q be a path from c to z in G that walks the edge (c, q) , then walks a shortest path from q to r in G_I^v , and finally walks the edge (r, z) . The length of this path is precisely $|Q| = 1 + (\ell - 1) + 1 = \ell + 1$, where ℓ is the number of layers in G_I . Now, consider any alternate path Q' from c to z in G . Note that all paths from c to z must include a path from the first layer of some copy of G_I to the last layer of that copy, in order to reach z in the last layer of G_O . This portion of the path has length $\ell - 1$. If Q' is as short as Q , i.e. $|Q'| \leq 1 + \ell$, then Q' may contain at most two other edges. In particular, these must be the first (incident on c) and last (incident on z) edge of the path, which are not contained in any copy of G_I .

Note that there is a unique path of length 2 from c to z in G_O , which uses the node v as its midpoint. Thus, the first (resp. last) edge in Q' necessarily connects c (resp. z) to nodes in G_I^v . By construction, the only edges in G that do so are (c, q) and (r, z) . Thus, any shortest path from c to z in G has the form described in Lemma 5: it walks the edge (c, q) , then it walks a shortest path from q to r in G_I^v , and finally it walks the edge (r, z) . \square

We may then prove Claim 7 via an identical proof to the one used in the weighted setting. Finally, we have this analogue of Claim 8:

Claim 9. *Any distance preserver H of G, P_O must keep*

$$\Omega\left(I^{2d/(d^2+1)} \left(\frac{n}{2^{\Theta(\sqrt{\log n \cdot \log \log n})}}\right)^{d(d-1)/(d^2+1)}\right)$$

edges in each inner graph G_I^v .

Proof. Identical to the proof of Claim 8. \square

Thus, we have a graph on $\Theta(nI)$ nodes, a set S of $s = \Theta(n)$ nodes (the first layer together with the last layer of G), and any preserver of G, S contains a preserver of G, P_O and thus has

$$\begin{aligned} & \Omega \left(I^{2d/(d^2+1)} n^{1+d(d-1)/(d^2+1)} \left(\frac{1}{2^{\Theta(\sqrt{\log n \cdot \log \log n})}} \right)^{d(d-1)/(d^2+1)} \right) \\ &= \Omega \left(I^{2d/(d^2+1)} n^{1+d(d-1)/(d^2+1)} \left(\frac{1}{2^{\Theta(\sqrt{\log n \cdot \log \log n})}} \right) \right) \end{aligned}$$

edges. We may rephrase this edge bound as

$$\Omega \left((nI)^{2d/(d^2+1)} s^{(2d-1)(d-1)/(d^2+1)} \left(\frac{1}{2^{\Theta(\sqrt{\log n \cdot \log \log n})}} \right) \right)$$

and since G has $\Theta(nI)$ nodes, this implies Theorem 4.

5 Acknowledgements

I am grateful to Amir Abboud for many useful discussions that have led to some of the central ideas in this paper. I am grateful to Noga Alon for suggesting some of the connections between distance compression and combinatoric graph theory covered by Theorem 2. I thank my advisor Virginia Vassilevska Williams for advice and technical discussions. Finally, I thank an anonymous reviewer for many very helpful comments and corrections that have greatly improved the current presentation of this work.

References

- [1] Amir Abboud and Greg Bodwin. Error Amplification for Pairwise Spanner Lower Bounds. In *Proc. of 27th SODA*, 2016.
- [2] Amir Abboud and Greg Bodwin. The $4/3$ Additive Spanner Exponent is Tight. In *Proc. of 48th STOC*, pages 351–361, 2016.
- [3] Donald Aingworth, Chandra Chekuri, and Rajeev Motwani. Fast Estimation of Diameter and Shortest Paths (without Matrix Multiplication). In *Proc. of 7th SODA.*, pages 547–553, 1996.
- [4] N. Alon, A. Moitra, and B. Sudakov. Nearly Complete Graphs Decomposable into Large Induced Matchings and their Applications. In *Proc. of 44th STOC*, pages 1079–1089, 2012.
- [5] Stephen Alstrup, Søren Dahlgaard, Mathias Bæk Tejs Knudsen, and Ely Porat. Sublinear distance labeling for sparse graphs. *preprint, arXiv:1507.02618*, 2015.
- [6] I Althöfer, G Das, D Dobkin, D Joseph, and J Soares. On sparse spanners of weighted graphs. *Discrete & Computational Geometry*, 9:81–100, 1993.
- [7] Baruch Awerbuch. Complexity of network synchronization. *Journal of the ACM*, pages 32,804–823, 1985.
- [8] Surender Baswana, Telikepalli Kavitha, Kurt Mehlhorn, and Seth Pettie. Additive spanners and (α, β) -spanners. *ACM Transactions on Algorithms*, 7(1):5, 2010.

- [9] Felix A Behrend. On sets of integers which contain no three terms in arithmetical progression. *Proceedings of the National Academy of Sciences of the United States of America*, 32(12):331, 1946.
- [10] D. Biló, F. Grandoni, L. Gualà, S. Leucci, and G. Proietti. Improved purely additive fault tolerant spanners. In *ESA*, pages 167–178, 2015.
- [11] Y. Birk, N. Linial, and R. Meshulam. On the Uniform-traffic Capacity of Single-hop Interconnections Employing Shared Directional Multichannels. *IEEE Transactions on Information Theory*, pages 186–191, 1993.
- [12] Greg Bodwin and Virginia Vassilevska Williams. Very Sparse Additive Spanners and Emulators. In *Proc. of 6th ITCS*, pages 377–382, 2015.
- [13] Greg Bodwin and Virginia Vassilevska Williams. Better Distance Preservers and Additive Spanners. In *Proc. of 27th SODA*, 2016.
- [14] Béla Bollobás, Don Coppersmith, and Michael Elkin. Sparse distance preservers and additive spanners. In *Proc. of 14th SODA*, pages 414–423, 2003.
- [15] G. Braunschvig, S. Chechik, D. Peleg, and A. Sealfon. Fault tolerant additive and (μ, α) -spanners. *Theoretical Computer Science*, pages 94–100, 2015.
- [16] Shiri Chechik. New Additive Spanners. In *Proc. of 24th SODA*, pages 498–512, 2013.
- [17] Shiri Chechik, Michael Langberg, David Peleg, and Liam Roditty. Fault tolerant spanners for general graphs. *SIAM Journal on Computing*, 39(7):3403–3423, 2010.
- [18] D. Conlon and J. Fox. Graph Removal Lemmas. *Surveys in Combinatorics*, pages 1–50, 2013.
- [19] Don Coppersmith and Michael Elkin. Sparse Sourcewise and Pairwise Distance Preservers. *SIAM Journal on Discrete Mathematics*, pages 463–501, 2006.
- [20] Marek Cygan, Fabrizio Grandoni, and Telikepalli Kavitha. On Pairwise Spanners. In *Proc. of 30th STACS*, pages 209–220, 2013.
- [21] M. Dinitz and R. Krauthgamer. Fault-tolerant spanners: better and simpler. In *PODC*, pages 169–178, 2011.
- [22] M Elkin. An improved construction of progression-free sets. *Israeli Journal of Math* 184, pages 98–128, 2011.
- [23] Michael Elkin. Computing almost shortest paths. *ACM Trans. Algorithms*, pages 1(2):283–323, 2005.
- [24] Michael Elkin, Arnold Filtser, and Ofer Neiman. Terminal embeddings. In *Proc. of APPROX-RANDOM*, pages 242–264, 2015.
- [25] Michael Elkin and David Peleg. $(1 + \epsilon, \beta)$ -spanner constructions for general graphs. *SIAM J. Comput.*, 33(3):608–631, 2004.
- [26] Michael Elkin and Seth Pettie. A Linear Size Logarithmic Stretch Path-Reporting Distance Oracle for General Graphs. In *Proc. of SODA*, 2015.

- [27] Paul Erdős. Extremal problems in graph theory. *Theory of graphs and its applications*, pages 29–36, 1964.
- [28] E. Fischer, I. Newmann, S. Raskhodnikova, R. Rubinfeld, and A. Samorodnitsky. Monotonicity Testing over General Poset Domains. In *Proc. of 34th STOC*, pages 474–483, 2002.
- [29] Jacob Fox. A new proof of the graph removal lemma. *Annals of Mathematics* 174, pages 561–579, 2011.
- [30] Telikepalli Kavitha. New Pairwise Spanners. In *Proc. of 32nd STACS*, pages 513–526, 2015.
- [31] Telikepalli Kavitha and Nithin Varma. Small Stretch Pairwise Spanners. In *Proc. of 40th ICALP*, pages 601–612, 2013.
- [32] Merav Parter. Bypassing Erdős’ Girth Conjecture: Hybrid Stretch and Sourcewise Spanners. In *Proc. of 41st ICALP*, pages 608–619, 2014.
- [33] Merav Parter. Vertex fault tolerant additive spanners. In *Distributed Computing*, pages 167–181, 2014.
- [34] Merav Parter. Dual failure resilient bfs structure. In *PODC*, pages 481–490, 2015.
- [35] Merav Parter and David Peleg. Sparse fault-tolerant bfs trees. In *ESA*, pages 779–790, 2013.
- [36] David Peleg and Alejandro Schaffer. Graph spanners. *Journal of Graph Theory*, pages 13:99–116, 1989.
- [37] Seth Pettie. Low Distortion Spanners. In *Proc. of 34th ICALP*, pages 78–89, 2007.
- [38] Seth Pettie. Low distortion spanners. *ACM Transactions on Algorithms*, 6(1), 2009.
- [39] I. Z. Ruzsa and E. Szemerédi. Triple systems with no six points carrying three triangles. *Coll. Math. Soc. J. Bolyai* 18, II:989–945, 1976.
- [40] Mikkel Thorup and Uri Zwick. Spanners and emulators with sublinear distance errors. In *Proc. of 17th SODA*, pages 802–809, 2006.
- [41] David P. Woodruff. Lower bounds for additive spanners, emulators, and more. In *Proc. of 47th FOCS*, pages 389–398, 2006.